Journal of Biomedical Informatics 46 (2013) 765–773

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Special Communication

Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine

Carol Friedman [a,*], Thomas C. Rindflesch [b], Milton Corn [b]

[a] Department of Biomedical Informatics, Columbia University, United States
[b] National Library of Medicine, Division of National Institutes of Health, United States

ARTICLE INFO

Article history:
Received 29 April 2013
Accepted 7 June 2013
Available online 25 June 2013

Keywords:
Natural language processing
Biomedical language processing

ABSTRACT

Natural language processing (NLP) is crucial for advancing healthcare because it is needed to transform relevant information locked in text into structured data that can be used by computer processes aimed at improving patient care and advancing medicine. In light of the importance of NLP to health, the National Library of Medicine (NLM) recently sponsored a workshop to review the state of the art in NLP focusing on text in English, both in biomedicine and in the general language domain. Specific goals of the NLM-sponsored workshop were to identify the current state of the art, grand challenges and specific roadblocks, and to identify effective use and best practices. This paper reports on the main outcomes of the workshop, including an overview of the state of the art, strategies for advancing the field, and obstacles that need to be addressed, resulting in recommendations for a research agenda intended to advance the field.

1. Introduction

There is an explosion of electronic information concerning health and health care, including publications, electronic health records, and the Web. Harnessing that information, most of which is in textual form, is critical for all aspects of health care: it is needed to drive innovation in research aimed at advancing health as well as to drive improvements in quality and in reducing costs. Natural language processing (NLP) is essential because it is needed to transform relevant information locked in text into structured data that it can be used by computer processes. In light of the importance of NLP to health, the National Library of Medicine (NLM) sponsored a two day workshop on April 24 and April 25, 2012 to review the state of the art in NLP focusing on text in English, both in biomedicine and in the general language domain. We invited researchers who are prominent in NLP in the biomedical domain as well as in the general domain to discuss strategies that show promise in facilitating significant progress (in robustness, generality, and accuracy) in the field, and to also define challenges that should be addressed. The workshop was held in conjunction with sessions hosted by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) that addressed the promise and application of NLP in clinical decision support, and the public was invited to participate as well.

Specific goals of the NLM-sponsored workshop were to identify the current state of the art, grand challenges and specific roadblocks, and to identify effective use and best practices. We sought recommendations for a research agenda intended to achieve progress in the field with a principled approach that:

- robustly generalizes to accommodate a range of linguistic expressions,
- exploits both the structure of language and of ontology as the basis for deeper understanding of phenomena underlying effective NLP,
- combines symbolic with statistical methods in an effort to balance linguistic theory and linguistic features occurring in text.

Drs. Carol Friedman and Thomas Rindflesch served as workshop co-chairs. The format of the workshop was based on invited speakers who were well-known in the field, and consisted of overview speakers, panel presentations, and breakout groups, which were organized under three themes: combining statistics and linguistic structure, focusing on linguistic structure, and applications in biomedical research.

Dr. Donald A.B. Lindberg, Director of the NLM, and Dr. Milton Corn, Deputy Director for Research and Education opened the

* Corresponding author. Fax: +1 212 305 3302.
E-mail address: friedman@dbmi.columbia.edu (C. Friedman).

1532-0464
http://dx.doi.org/10.1016/j.jbi.2013.06.004

workshop with welcoming remarks. Dr. Lindberg, who has been an advocate and supporter of NLP for many years, commented that he greatly admires the field and hopes it will strive for natural language understanding, rather than just processing. He noted that progress is being made in developing effective tools based on NLP, such as automatic indexing at the NLM [1], but the more ambitious goal of modeling human cognition remains elusive. He emphasized that it is important to maintain momentum and that a major benefit of effective NLP would be to provide convenient access to, as well as some understanding of, the vast amount of medical knowledge currently sequestered in clinical text.

Dr. Corn, who headed the workshop organizing committee, pointed out that NLP is crucial for making computers maximally valuable generally, and for biomedical research in particular, and deserves considerable attention and support. He stated that the NLM is strongly interested in finding out what the broad NLP research community considers to be important directions for research as well as challenges that should be addressed, and that this information will provide valuable guidance for NLM's efforts in helping to further the field.

Carol Friedman was the keynote speaker. She provided an overview of NLP work in the biomedical field, summarized the milestones, and discussed the need for development of new NLP paradigms involving integration of statistics, linguistic knowledge, and domain knowledge. She stated that to energize the field, there is a need for development of clinically striking NLP applications that can be widely used. There were some challenges and future directions that were shared among the presenters as well as some unique ones. The most common obstacle for NLP that was mentioned throughout the meeting was the inaccessibility of large scale de-identified clinical corpora, which is needed for training and evaluating NLP systems. Additionally, two future directions proposed by many of the speakers as well as the participants was a need to develop methods that utilize and integrate knowledge and statistical methods, and also a need to move towards natural language understanding.

The first two speakers presented overviews of the state of the art. Christopher Manning presented statistics-focused approaches. He pointed out that although supervised learning for NLP has generated good results, it requires costly manual annotation, and future directions should explore automated utilization of online knowledge for training. Two of his other suggestions for future work were to explore improved understanding and to adapt the current approach to NLP, where the overall task is broken up into a sequence of independent processing modules, to a model where modules can interact. Sergei Nirenberg presented an overview of linguistics-based methods, stating that most systems today were hybrid. He felt that the grand challenge for future research would be in the development of a system that can communicate with and aid humans. He noted that different methods would be needed for human computer interaction than for information extraction, which is currently the focus in the field. He also felt that progress will depend on integration of ontological, linguistic, and world knowledge.

The first panel discussed the integration of linguistic knowledge and statistical methods. Dan Moldovan suggested several options for combining the two methods that involved voting or using features from one approach and integrating them with the other approach. Philip Resnik mentioned the lack of large scale clinical text, which has impeded development of statistical models. Interestingly, he pointed out that increased adoption of Electronic Medical Record systems may present an even bigger challenge for NLP than the lack of data because these systems curtail the narrative description of the clinical encounter and encourage entry of discrete data points. Jun'ichi Tsujii described the GENIA text annotation system, which has been used to develop NLP systems in the biomolecular domain. It consists of linguistic annotations based on grammatical information and language-independent knowledge-based mark-up consisting of concepts, entities, and relations in the domain.

The second panel discussed methods focused on linguistics and knowledge. Chitta Baral stated that NLP should move from a focus on extraction to understanding, which requires an amalgamation of facts, domain knowledge, general knowledge, and reasoning. Since knowledge is described in text, it would be ideal to extract it and store it in structures which can be accessed for subsequent computational processes. Lynette Hirschman discussed the importance of overcoming the barriers to obtaining clinical data in order to enable development of NLP systems that can scale. She mentioned that effective de-identification software and data-use agreements are critical to removing some of the barriers to clinical NLP, and that an important next step would be identification and development of realistic clinical tasks. James Pustejovsky focused on the modeling of temporal information inherent in text, which is critical for many reasoning tasks. Linguistic modeling of time involves determining the time of the events mentioned in text, and then enabling the ordering of the events.

The third panel discussed NLP applications that utilize linguistic explicit knowledge. Alan Aronson presented the work he spearheaded at the NLM, consisting of MetaMap [2], which is widely used by the research community, and employs linguistic knowledge to map text to UMLS [3] codes. Two other efforts he discussed were the Medical Text Indexer [1], which is used in production at the NLM to recommend MeSH terms to expert indexers, and the recently developed Gene Index Assistant. Kevin Cohen talked about translational NLP in the biomedical domain (BioNLP) and a need to include the vast amount of biomedical knowledge and ontologies in NLP as well as the potential for handling the very complex verb-dominated biomolecular language utilizing sublanguage theory. Thomas Rindflesch discussed SemRep [4], which he was responsible for developing at the NLM. SemRep, which is based on linguistic symbolic principles, is widely used to extract predications needed for biomolecular text mining of the literature as well as clinical research. He described the broad uses of SemRep, demonstrating its generalizability and portability.

The remainder of the workshop consisted of the breakout groups followed by summarization by the moderators and then a discussion of the overall issues by all the participants. An exciting proposal made by John Hurdle from the University of Utah has the potential to alleviate the bottleneck of access to clinical reports, and led to a subsequent flurry of emails and to joint work involving surveying consumers at different sites. His idea was an effort where patients could readily donate their de-identified clinical reports to a repository for research purposes. This would involve preparatory work and a group effort.

2. Overview presentations

2.1. Keynote

Carol Friedman (Columbia University)

Research in biomedical NLP has increased enormously in the last 30 years. It has become a prominent activity because of a critical need to harness the explosive amount of information in text concerned with biomedical research, clinical care, as well as health-related information on the Internet, and of a need to use the information for applications concerned with many facets of health. Fig. 1, which presents the number of publications on NLP in MEDLINE, shows a big increase starting in the 1990s.

In the clinical domain critical applications, such as decision support, cohort identification, patient management, resource manage-

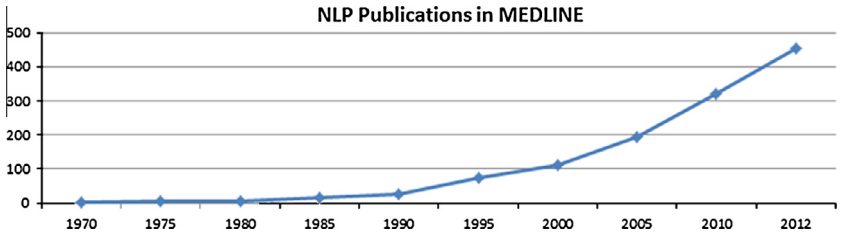

Fig. 1. Number of NLP publications in MEDLINE.

ment, question answering, knowledge acquisition, research, and discovery require NLP in order to structure the information in text to render more reliable access of the information. In the broader biomedical domain, critical applications such as information retrieval, database curation, knowledge discovery, knowledge acquisition and management, and tailoring information for consumers also need NLP.

The start of clinical NLP began in the 1960s. Pioneering work by Naomi Sager in the 1970s and 1980s, based on language theories of Zellig Harris [5–7], demonstrated that it was actually feasible to structure clinical information [8] occurring in text. Starting in the late 1980s, other early NLP systems also demonstrated that NLP was feasible in the clinical domain [9–16], and that NLP could actually be used to improve clinical care [17–19]. Substantial resources for NLP became available in the 1980s and 1990s. For example, structured domain knowledge was provided by SNOMED [20], a large clinical healthcare terminology. The Unified Medical Language System (UMLS) [3], an integrated compendium of more than 150 terminologies in the biomedical sciences, was a major resource for NLP. One component of the UMLS, the Specialist Lexicon [10], contains syntactic information for a significant part of biomedical language. PubMed (http://www.ncbi.nlm.nih.gov/pubmed) is a bibliographic database comprising more than 21 million citations from the biomedical research literature; entries are annotated with Medical Subject Heading (MeSH) terms (http://www.nlm.nih.gov/mesh). PubMedCentral (http://www.ncbi.nlm.nih.gov/pmc) is a repository of full-text journal articles in the domain. NLP research in the biomolecular domain started in the late 1990s to capture, organize, and connect information from journal articles [21–25]. These efforts were followed by community research efforts that included creation of annotated corpora from journal articles, which were made available to the community [26,27] and community-wide NLP challenges [28], which initially involved the processing of journal articles, and subsequently involved processing of clinical reports, although on a much smaller scale [29–33]. In the early 2000s open source NLP tools in the biomedical domain also became available, which can now be registered and accessed online via the Orbit Project (http://orbit.nlm.nih.gov).

NLP has many different aspects, as illustrated in Fig. 2. The left side of the figure shows that sample corpora, a domain model, and domain and linguistic knowledge are needed to develop NLP systems. A primary challenge and bottleneck for progress in our field is limited access to clinical notes, which is due to patient confidentiality issues. In general, only researchers affiliated with a medical center can access clinical notes with Institutional Review Board approval, but these notes cannot be shared with the external NLP community unless they are de-identified and data use agreements are in place, but even then, many institutions are hesitant to allow clinical notes to be disseminated for community-wide research purposes even if they are de-identified. Once a system is developed, evaluations are important for progress so that the different systems can be compared, but this also requires sharing of clinical notes. Evaluations are needed because they enable us to learn the factors involved in obtaining the results, such as what NLP components were critical to the task, how much domain knowledge and reasoning were involved in addition to NLP, what were the primary causes of errors, and how generalizable are the methods. The right hand side of Fig. 2 shows the operational aspects of NLP, including methods, tools, systems, and applications. A big challenge is NLP methodology itself. We need to keep exploring new as well as incremental NLP methods. Linguistic trends have swung from empirical corpus-based methods to symbolic rule-based methods, back to the current approach, which focuses on statistical corpus-based methods. Each method has advantages and disadvantages, and research is needed in development of synergistic NLP models that integrate linguistic rules, domain knowledge, and statistical methods. NLP by itself is not the goal, but a means for enabling other computational processes. Therefore it is critical that we experiment with a variety of clinical applications with high visibility to help further the field.

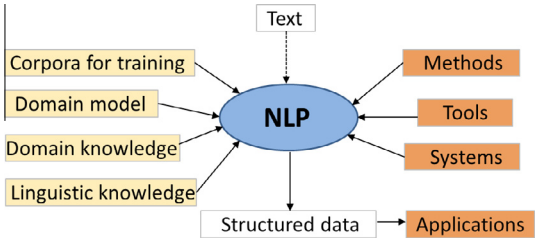

Fig. 2. Different aspects of NLP.

In conclusion, using considerable available resources, progress has been made in NLP, especially in recognizing entities in text (e.g. persons, procedures, drugs, diseases, body substances, genes, etc.) effectively enough for practical use. In order to approach deeper levels of understanding, continuing efforts must be made in recognizing relations among these entities (e.g. drug X causes adverse event Y or drug X treats adverse event Y). In addition to such relations, interpreting higher levels of meaning, such as beliefs, opinions, and intentions must be addressed.

2.2. Statistics-focused approaches

Christopher Manning (Stanford University)

The use of statistics (probabilities) in NLP is motivated by a desire to compensate for the difficulty of creating a grammar that adequately covers all language and that provides for reasoning and uncertainty. Using probabilities helps a system to make good guesses, and complements, rather than supplements, good linguistic representation. Probabilities associated with textual terms that are calculated over large amounts of data can often provide very useful results. The primary NLP successes were based on supervised learning which requires annotation of data sets and a classifier to reproduce the annotations. These statistical methods do not remove linguistic knowledge but externalize it via the annotations. Statistical methods have been applied to machine translation, syntactic parsing (both phrase structure and dependencies), named

entity recognition, relation extraction, and co-reference resolution. Several important directions for research were suggested, and a summary is shown in Fig. 3. One involved finding alternatives to supervised learning because of its limitations: annotation is costly and results deteriorate when moving to different domains; therefore the amount of training data will always be relatively small compared to the amount of text. In an attempt to alleviate the onus of annotated text needed by the supervised learning paradigm, there is a need to build systems that require less supervision, which involves exploring the use of information in other sources, such as information in databases, ontologies, or in Wikipedia, where the challenge would likely be to overcome noise in the use of those data for training. Another direction worth exploring involves improving on the pipeline approach toward NLP. For example, the typical system performs a series of independent tasks aimed at regularizing text, such as tokenization, part of speech tagging, named-entity recognition, parsing, and extraction. However, joint inferencing among modules may be a better paradigm. Another important research direction should be aimed at improved understanding of meaning beyond interpretation of individual words. Such extension includes first understanding relations among words in text (e.g. does drug X treat, or cause, adverse event Y), and then determining whether relations asserted in text are true (e.g. is drug X actually effective in treating Y). This level of understanding is needed to provide a foundation for tasks involving knowledge and reasoning. Therefore, much work remains to achieve the language capability of science fiction robots. Fig. 3, which is the conclusion slide of Manning's talk, summarizes his main points.

2.3. Linguistics-focused approaches

2.3.1. Sergei Nirenburg (University of Maryland, Baltimore County)

An overview of linguistics-based NLP systems in biomedicine illustrates that most recent systems are hybrid, combining both statistical- and knowledge-based methods, and that all are ultimately linguistics based. Rule-based approaches use experts whose knowledge is explicitly represented by rules, whereas statistical approaches utilize linguistic features as well as other features, and require annotated corpora for training and testing. Much of the NLP research in the biomedical domain has been focused on information extraction (IE), which can also be used as input to other technologies, such as clinical decision support. In addition to supporting IE, some NLP systems are being developed to facilitate medical training and decision making. NLP capabilities required for human–computer interaction (HCI) and for decision-making modules are different from those needed to support IE. The ultimate criterion of success for any system is user acceptance. Several user-acceptance studies demonstrate that currently available systems do not approach the capabilities of HAL (from the movie 2001), a robot that understood and responded to user requests. Fig. 4 is a slide from Nirenberg's presentation which explains why we do not have the understanding capabilities of HAL. The grand challenge for future research is to build a system that can meaningfully communicate with humans. This would require customizability, non-interference with the user's workflow, and ability to provide cognitive support. Significant progress will crucially depend on the system having available considerable ontological, linguistic, and world knowledge as well as the ability to model beliefs of its human interlocutors and both human and artificial team members, and on the ability of the system to integrate those components.

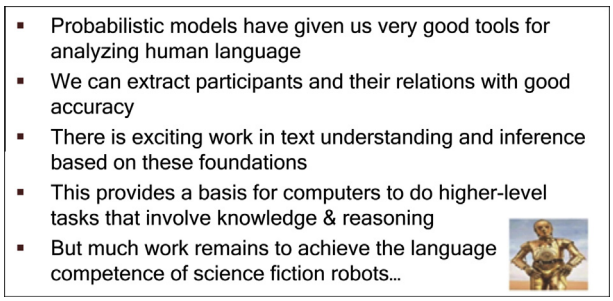

Fig. 3. Conclusion slide from Chris Manning's presentation.

3. Panel presentations

3.1. Panel 1: Combining statistics and linguistic structure

3.1.1. Dan Moldovan (Human Language Technology Research Institute, University of Texas at Dallas): "NLP for the medical domain"

A suite of tools developed at the Human Language Technology Research Institute includes both statistical and rule-based approaches [34,35]. The tools follow the pipeline approach to generate a structured interpretation of meaning from text, where one module feeds into the next. The functions of the modules range from low level NLP to deep semantic analysis of text and perform the following tasks: tokenization, part-of-speech tagging, sentence boundary detection, named-entity recognition, concept tagging, syntactic parsing, word sense disambiguation, context detection, semantic parsing, co-reference resolution, event extraction, and semantic calculus. Notably, a semantic parser combines linguistic analysis, domain knowledge, and machine-learning classification. The semantic parser is a core module in an ontology building tool that automatically builds domain ontologies from domain documents and has an interface that allows visualization and editing of ontologies. Additionally, the knowledge representation is hierarchical, allowing for various levels of representation ranging from lexical concepts to macro-events. Research indicates that the "semantic calculus" can extract additional high-level relations and can help with the rapid customization of semantic relations extracted from text. Statistical and knowledge-based semantic approaches can be combined for each module in several ways: by voting, by using features from the knowledge-based approach for statistical methods, or by using statistical methods to filter results, and then use knowledge-based methods.

3.1.2. Philip Resnik (University of Maryland): "Clinical NLP and the data dilemma"

State of the art NLP is driven by statistical methods dependent on large-scale data analysis, with crucial contributions from linguistic knowledge and knowledge of the application domain. Progress on statistical methods in clinical NLP faces an enormous challenge, though, because the HIPAA Privacy Rule (HHS, 2013) and institutional concerns make it difficult for the broad community of NLP researchers and developers to obtain access to relevant data. Efforts in clinical NLP are based on orders of magnitude less data than other NLP efforts [33] or restricted to researchers connected with a hospital or other healthcare organization, or both. As a result, despite its importance, clinical NLP has not developed the energy and pace of progress seen in other application domains, e.g. the analysis of social media data [36]. An even more fundamental challenge is the effect that EMR systems could have on natural clinical language as they become more widely adopted. When EMR systems replace the spoken or text description of an encounter with drop down menus or point and click tools for discrete data entry, they undermine the capture of the full clinical narrative – which is not just an unconnected set of data points, but a rich and nuanced record of the patient's condition, the temporal

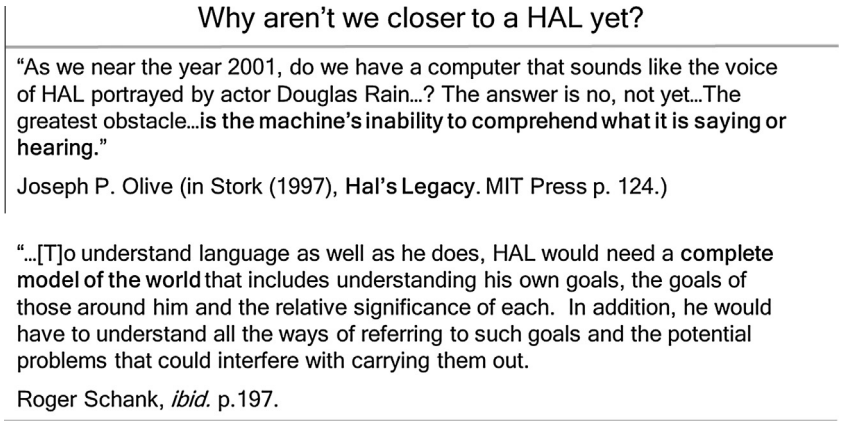

Fig. 4. Slide from Sergei Nirenberg's presentation.

sequence of events, and the clinician's thought process [37]. More effort is needed to preserve naturally occurring clinical language and to make it availability to the research community.

3.1.3. Jun'ichi Tsujii (Microsoft Research Asia): "Information extraction, parsers, and ontology"

GENIA is a text annotation system for extracting molecular information that occurs in text where the text types are abstracts and complete journal articles, and where the objective is to represent and record the relevant information in a structured form [26,38]. The system incorporates the language domain, which involves linguistic expressions, and the knowledge domain, which is motivated independently of language and involves concepts and relationships in the domain. GENIA uses the Gene Ontology (GO) [39] and MeSH for the named entity ontology, and GO for the event ontology, whereas it has its own relation and meta-knowledge ontology. Text annotation involves mapping the text into structures represented in the knowledge domain, and consists of both linguistic (syntactic) and semantic annotations. Syntactic annotation includes part-of-speech tagging, syntactic parsing of phrases, and further analysis to obtain deep argument structure (e.g. identify predicates and their arguments) and to identify discourse elements whereas semantic annotation involves named entity (NE), event, and relation identification, coreference, and meta-knowledge. To achieve better performance when moving to another domain or task, the linguistic component of the system must be adapted. Complementing a statistically-based system with semantic and ontological knowledge will help to overcome limitations of statistical-based systems. The feasibility of NLP is highly dependent on text types, domains, and tasks. GENIA's domain is molecular biology, where the text types are full journal articles or abstracts. Another domain is the clinical domain where the text types are patient records, such as discharge summaries. The language of this domain is different in that verbs are not central, and many inferences are needed to understand the text [40].

3.2. Panel 2: Linguistics-based methods

3.2.1. Chitta Baral (Arizona State University): "From NLP to natural language understanding for medical decision making"

Translating natural language text to a formal logic and reasoning with that logic constitutes a step toward natural language understanding. More specifically, understanding requires facts, and domain and general knowledge, all of which may occur in text and therefore should be extracted. It is important to develop generic approaches to fact extraction. One approach would involve minimizing processing by storing annotated semantic information and parse trees in a database structure where the parse tree can be queried using a specialized parse tree query language [41]. In addition to NLP techniques to extract relations, the method includes building reasoning chains to move NLP applications towards decision making. Goals include extracting information such as knowledge on protein–protein interactions and pharmacokinetics to make hypotheses regarding drug–drug interactions [42].

3.2.2. Lynette Hirshman (The MITRE Corporation): "Scaling the data wall"

The application of NLP to clinical records (clinical NLP) holds out great promise of making computable the rich information contained in electronic health records – provided that we can both "scale the data wall" and scale the data. Scaling the data wall will make it possible to assemble corpora of (de-identified) medical records for research; scaling the data will enable the application of clinical NLP to solve pressing problems, such as selection of patient cohorts for research or elucidating genotype-phenotype relations. NLP has been an active research field since the 1960s, with enormous progress, driven by development of shared corpora, such as the Penn Treebank [43] and challenge evaluations, such as the Message Understanding Conferences [44,45]. However, application of NLP to the clinical domain has been more challenging, despite the existence of large-scale terminologies and ontologies, as well as access to electronic text from biomedical journals. A major stumbling block has been the difficulty of obtaining sharable corpora of medical records; lack of such shared data sets has made it impossible to compare performance of NLP systems for clinical tasks. Fortunately, over the past few years, there has been significant progress; there are now several medical record corpora that have been made available under limited data use agreements, and there are improved tools to facilitate de-identification of medical record corpora [46,47], as well as shared annotation tools to support annotation with both medical and linguistic constructs. This has laid the ground work for challenge evaluations such as the series of i2b2 challenge tasks [33]. Creation of these resources has been an important first step [36]; however, to demonstrate the promise of clinical NLP, it is now critical to take the next step, namely applying these techniques to realistic clinical tasks, in order to make accessible the rich information contained in medical narrative. This will require cost-effective, scalable methods of de-identification and preparation of training data. New multi-site projects such as eMERGE [48] and open source modules, such as cTAKES [49] will help to further scale the data wall that impedes corpora sharing and NLP development.

3.2.3. James Pustejovsky (Brandeis University): "Reasoning about temporal and event information in clinical texts"

Temporal extraction is a very challenging, intricate, and critical NLP task. Because clinical text contains references to past, current, future, and planned events, dealing with time is important for clinical decision making. An objective for dealing with time consists of detecting relevant events in text, anchoring the events to temporal expressions, and then ordering the event temporally. Some current crude methods for handling time use the time of document creation to link all events in a document to that time point, or find an event and local temporal expression and link the event to that expression. However, time should be modeled more precisely, which is the goal of TimeML [50,51]. It involves modeling events relative to time, such as dates and durations, and modeling temporal relations, such as *before* and *during*. Linguistic modeling of temporal information is a critical component in the development of expressive specification languages for annotation markup as elements of the specification are the foundation for feature identification underpinning statistics-based NLP.

3.3. Panel 3: Linguistics in biomedical applications

3.3.1. Alan Aronson (National Library of Medicine): "NLM Indexing Initiative tools for NLP: MetaMap and the Medical Text Indexer"

MetaMap identifies Unified Medical Language System (UMLS) concepts in text based on linguistic principles [2]. It uses a minimal commitment parser, lexicon, and part-of-speech tagger, all developed at the NLM. It then retrieves candidate terms from the UMLS Metathesaurus, and scores the terms based on an evaluation function. It also includes a word-sense disambiguation facility, recently enhanced with a statistical context-sensitive method. MetaMap underpins the Medical Text Indexer (MTI), which summarizes text using the Medical Subject Heading (MeSH) terminology [1]. MTI has been used in production since 2002 for indexing MEDLINE citations and Cataloging and History of Medicine records at the NLM. It processes the abstracts and then recommends MeSH terms, which are reviewed by experts who select, revise, and approve the terms. In February 2011, MTI became the first-line indexer (MTIFL) for a select number of journals, where it has historically performed well. The MTIFL indexing for these journals is only revised by an indexer. The Gene Indexing Assistant (GIA) is a new automated tool being developed to assist indexers in identifying and creating GeneRIFs (gene reference into function), which provide a mechanism that enhances the functional annotation of genes [52].

3.3.2. Kevin Cohen (University of Colorado): "Translational bioNLP"

The time is ripe for translational BioNLP—leveraging the results of fourteen years' worth of work on genomics journal articles to do better clinical NLP. One opening for doing this stems from the observation by Friedman et al. [40] that clinical text is noun-dominated, while biological journal articles are verb-dominated. It turns out that nominalizations—nouns derived from verbs—are prominent features in biological journal articles, where they exhibit highly complex and interesting patterns of syntactic argument realization [53]. Despite this complexity, nominalizations are tractable for NLP for two reasons. One is that we can take advantage of the observations that within a sublanguage many predicators have only limited possibilities for fillers of their argument slots [54] and that these fillers belong to specific semantic classes [55]. The other opening for doing translational NLP is that we have reached a point where biomedical background knowledge and ontologies are rich enough to leverage in NLP. Even small amounts of knowledge can improve performance. For example, Livingston and colleagues [56] were able to improve information extraction about gene activation events by 20 points of F-measure just with knowledge of whether specific genes had Gene Ontology annotations of *catalytic activity* or *receptor activity* [55]. More complex resources from the NLM make more complicated analyses possible. For example, we can differentiate between constructions like *phenobarbital treatment* (treatment *with* phenobarbital) and *cancer treatment* (treatment *of* cancer) using the *Relations* field in the UMLS. This brings us back to how to use Friedman's et al. [40] observation about the noun-domination of clinical text. Domain knowledge allows the recovery of missing arguments, and identification of non-absent arguments. For relational nouns in particular, we can infer metonymically implied arguments that are necessary for the full representation of meaning [54]. Thus, bringing together work on nominalization in biomedical journal articles with the currently available biomedical background knowledge and ontologies makes this a fruitful time for translational BioNLP.

3.3.3. Thomas Rindflesch (National Library of Medicine): "Knowledge-based NLP in biomedicine"

SemRep [4] is a symbolic NLP system that currently identifies semantic propositions in MEDLINE citations and is being extended to interpret extra-propositional meaning, such as speculations, opinions, evidence, and attitudes. Based on generalizations about English syntax and structured domain knowledge from the UMLS, it balances linguistic insight with implementation expediency to identify 26 core predicates limited by domain. The system has been used to extract more than 60 million semantic predications from all of MEDLINE (more than 21 million titles and abstracts). These predications are stored in a MySQL database made available to the research community [57]. While SemRep originally applied to clinical research as well as molecular biology, it is being extended to epidemic preparedness, climate change and health, health promotion, and biomedical knowledge processing. SemRep supports Semantic MEDLINE, a Web application for managing the results of PubMed searches [58,59]. This application is being exploited in research on literature-based discovery [60,61], portfolio analysis for NIH grant applications, and for finding the literature to support clinical practice guideline development [62].

4. Reports from the breakout groups

Attendees were invited to participate in a breakout group focused on one of the three workshop themes. Each group was moderated by researchers in the field, who reported discussion results to the workshop co-chairs. Breakout group moderators were Wendy Chapman (University of California, San Diego), Marcelo Fiszman (NLM), Graciela Gonzalez (Arizona State University), Halil Kilicoglu (NLM), Hongfang Liu (Mayo Clinic), and Serguei Pakhomov (University of Minnesota). The breakout groups facilitated the articulation of ideas from researchers with diverse backgrounds. Moderators encouraged suggestions for addressing confounds to achieving effective results. Although a range of opinions was expressed, there was consensus on several points.

4.1. Input text

A distinction was made between biomedical text sources, including colloquial texts (mailing lists, health-related Web sites and blogs, biomedical tweets), biomedical literature, and clinical text.

To facilitate more effective NLP, participants recommended that a classification of clinical reports in the EMR be devised that should include, but not be limited to:discharge summaries, radiology reports, pathology reports, clinical notes, nursing notes, and specialty notes. The kinds of knowledge (i.e. linguistic, ontological, pragmatic, etc.) needed to support high quality NLP should be described for each report type in the classification. This description should be

based on consideration of a report type compared to other types as well as report types considered longitudinally across the patient encounter. This description involves sublanguage modeling of report type and will facilitate both statistical and symbolic methods. In addition, since annotation is critical for NLP development, there should be a consensus on annotation.

A major challenge in the field is to make corpora widely available for researchers. Such corpora must be diverse to account for all text input categories and must accommodate a variety of use cases. It was agreed that a central repository of corpora representing a wide diversity of textual categories would be extremely valuable. The availability of clinical text corpora is particularly challenging due to confidentiality and de-identification requirements imposed by HIPAA restrictions and IRB regulations. Three suggestions to address these issues were made: to investigate the use of data from deceased patients, to take software to institutions that hold relevant text, or to institute a mechanism that would enable consumers to consent to donate their notes assuming that they would have de-identification software available. A subgroup formed, spearheaded by John Hurdle (University of Utah), as a result of this workshop to help further this effort.

4.2. Methodology

Discussion in the groups began with agreement that there are two major approaches to NLP, one, a symbolic method, gives priority to facts about linguistic structure and world knowledge; the second, a statistical method, relies on frequency of occurrence and distribution patterns of text tokens. It was noted that there is no such thing as a purely statistical method, since the underlying features are word-based. Linguistic representations and probabilities are not incompatible, and the most effective systems may combine the two approaches. There are known challenges inherent in both methods. Statistics-based systems need annotated text which cannot be used in other domains without influencing performance, and the output does not provide insight or the capability to correct errors. Symbolic systems require substantial expertise to develop the rules, and the complexity that is required is often a confounding element in development. In addition, there is a need for structured knowledge, both linguistic and ontological, which may not be readily available.

Other discussions concerning NLP methodology evolved around a variety of topics:

- The community must develop NLP systems with abilities to process ill-formed input and input from templates.
- NLP methods should be general while at the same time enabling development of specific tasks via tools that utilize general NLP methods.
- There needs to be more focus on social media text processing, such as Facebook, support groups, and Twitter feeds.
- The field requires evaluations concerning realistic applications where evaluation leads to a better understanding of the difficulty of the task and of error rates and types that are tolerable for different applications.
- The research community must push the envelope beyond NLP development and move towards the grand challenge of Natural Language Understanding.
- Research would benefit from a set of use cases for NLP.

4.3. Knowledge

There was broad agreement that more domain and linguistic knowledge is needed for significant advances in effective NLP, regardless of the approach, but that there is a significant gap between existing knowledge sources and their usefulness for NLP. Novel datasets that are not perfect but could be helpful should be exploited as much as possible. In addition to existing knowledge bases about medications and treatments, diseases and symptoms, and care plans, there is useful information for NLP in an often overlooked resource: legacy diagnostic artificial intelligence systems such as QMR [63], DxPlain [64], and ILIAD [65], but they may not be readily accessible for researchers. Knowledge extracted from online text, such as the biomedical literature (MEDLINE), encyclopedias (especially Wikipedia) and textbooks may be exploitable. Finally ontologies hold a central position as a knowledge resource underpinning NLP. The structure of ontologies representing classes of concepts and relationships in some domain is particularly relevant for semantic processing. In the biomedical domain there are existing ontologies and terminologies available in several components of the UMLS and from the National Center for Biomedical Ontology. More generally, WordNet [66] may be useful. Available ontologies are often sparse and may need to be enhanced and adapted with text-driven methods for NLP use.

5. Suggested research agenda

Based on substantial agreement among workshop participants, we make several recommendations to facilitate progress in the field:

- Methods that scale need to be developed, while utilizing and integrating linguistic and domain knowledge with statistical processing.
- The standard paradigm in NLP, a pipeline of independent modules, can be improved by allowing the modules to interact, with joint inferencing among modules being exploited.
- NLP needs to attain improved understanding of text, going beyond simple word meaning to interpreting relations among phrases and to advanced communication capabilities. Translating natural language text to a formal logic and reasoning with that logic constitutes a step toward this goal. Building reasoning chains with that logic can move NLP applications in that direction.

Regarding available resources, the most significant confounding element for clinical NLP is inaccessibility of large scale de-identified clinical corpora, which are needed for training and evaluation.

- An innovative proposal was for patients to donate their de-identified clinical reports to a repository for research purposes.

An additional issue is that statistical methods require annotation, which is costly, and results deteriorate when moving to different domains.

- Exploiting information in sources such as databases, ontologies, or Wikipedia to lessen the amount of supervision required might address this burden. Creatively exploiting existing knowledge sources can provide much-needed knowledge for NLP. In addition to clinically oriented knowledge bases, there is useful information for NLP in an often overlooked resource: legacy diagnostic artificial intelligence systems, or knowledge extracted from online text, such as the biomedical literature (MEDLINE), encyclopedias (especially Wikipedia) and textbooks, may also be exploitable.
- Finally, ontologies currently hold a central position as a knowledge resource underpinning NLP and should be exploited more vigorously.

6. Conclusion

In this article, we summarized the highlights of an NLM-sponsored workshop which recently reviewed the state of the art in NLP for English, both in biomedicine and in the general domain. Researchers prominent in the field met to discuss challenges and promising strategies for making significant progress in robustness, generality, and accuracy. Sessions with invited speakers were followed by break-out groups in which all those attending were invited to participate. A diverse group of very passionate researchers expressed a range of opinions and ideas, contributing to a robust and lively forum that provides direction for the continuing development of effective NLP applications. We were very gratified by the enthusiasm and acumen of all the participants, and thank them for their valuable contributions to the workshop. There are many avenues to be explored for NLP researchers, and we look forward to a productive and exciting future for NLP.

The workshop was videotaped and archived at http://www.videocast.nih.gov/ (search for "natural language processing"). Slides for all presentations are available at http://www.tech-res.com/NLPCDS.

Acknowledgment

We are grateful to Graciela Rosemblat and Dongwook Shin for assistance with the figures.

References

[1] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM indexing initiative's medical text indexer. Stud Health Technol Inform 2004;107(Pt 1):268–72.
[2] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229–36.
[3] Lindberg D, Humphreys B, McCray AT. The unified medical language system. Method Inform Med 1993;32:281–91.
[4] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003;36(6):462–77.
[5] Harris Z. Mathematical structures of language. NY: Wiley Interscience; 1968.
[6] Harris Z. A Grammar of English on mathematical principles. NY: Wiley & Sons; 1982.
[7] Harris Z. A theory of language and information: a mathematical approach. Oxford: Clarendon Press; 1991.
[8] Sager N, Friedman C, Lyman M, et al. Medical language processing: computer management of narrative data. Reading, MA: Addison-Wesley; 1987.
[9] Wingert F. Automated indexing of SNOMED statements into ICD. Method Inf Med 1987;26(3):93–8.
[10] McCray AT. Extending a natural language parser with UMLS knowledge. In: Proc Annu Symp Comput Appl Med Care; 1991. p. 194–198.
[11] Baud RH, Rassinoux AM, Scherrer JR. Natural language processing and semantical representation of medical texts. Method Inform Med 1992;31(2):117–25.
[12] Sager N, Lyman M, Tick LJ, Nhan NT, Bucknall CE. Natural language processing of asthma discharge summaries for the monitoring of patient care. In: Proc annu symp comput appl med care; 1993. p. 265–8.
[13] Friedman C, Alderson PO, Austin J, Cimino JJ, Johnson SB. A general natural language text processor for clinical radiology. JAMIA 1994;1(2):161–74.
[14] Haug P, Koehler S, Lau LM, Wang P, Rocha R, Huff S. A natural language understanding system combining syntactic and semantic techniques. In: Proc annu symp comput appl med care; 1994. p. 247–51.
[15] Moore GW, Berman JJ. Automatic SNOMED coding. In: Proc annu symp comput appl med care; 1994. p. 225–9.
[16] Berman JJ, Moore GW. SNOMED-encoded surgical pathology databases: a tool for epidemiologic investigation. Mod Pathol 1996;9(9):944–50.
[17] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports. Ann Int Med 1995;122(9):681–8.
[18] Fiszman M, Haug PJ. Using medical language processing to support real-time evaluation of pneumonia guidelines. In: Proc 2001 AMIA symp; 2000. p. 235–9.
[19] Aronsky D, Fiszman M, Chapman WW, Haug PJ. Combining decision support methodologies to diagnose pneumonia. In: Proc AMIA symp; 2001. p. 12–6.
[20] Cote RA, Robboy S. Progress in medical information management. Systematized nomenclature of medicine (SNOMED). Union Med Can 1980;109(9):1243–52.
[21] Rindflesch TC, Hunter L, Aronson AR. Mining molecular binding terminology from biomedical text. In: Proc AMIA symp; 1999. p. 127–31.
[22] Park JC, Kim HS, Kim JJ. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. Pac Symp Biocomput 2001:396–407.
[23] Tsujii J, Ohta T. Natural language processing for text mining in genome science. Tanpakushitsu Kakusan Koso 2001;46(Suppl. 16):2532–7.
[24] Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inform 2004;37(1):43–53.
[25] Koike A, Kobayashi Y, Takagi T. Kinase pathway database: an integrated protein-kinase and NLP-based protein-interaction resource. Genome Res 2003;13(6A):1231–43.
[26] Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus–semantically annotated corpus for bio-textmining. Bioinformatics 2003;19(Suppl. 1):i180–2.
[27] Vincze V, Szarvas G, Farkas R, Mora G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinform 2008;9(Suppl. 11):S9.
[28] Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinform 2005;6(Suppl. 1):S1.
[29] Pestian J, Brew C, Matykiewicz P, Hovermale D, Johnson N, Cohen K, et al. A shared task involving multi-label classification of clinical free text. In: Proc of ACL BioNLP; 2007.
[30] Uzuner O, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. Artif Intell Med 2008;42(1):13–35.
[31] Uzuner O, Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. J Am Med Inform Assoc 2009;16(1):109–15.
[32] Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenges on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2010;18(5):552–6.
[33] Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552–6.
[34] Blanco E, Moldovan D. Unsupervised learning of semantic relation composition. Portland, OR. In: Proc of the 49th annual meeting of the association for computional linguistics: human language technologies (ACT-HLT 2011); 2011.
[35] Moldovan D, Blanco E. Polaris: lymba's semantic parser. Istanbul, Turkey. In: Proc. of the 8th international conference on, language resources and evaluation (LREC'12); 2012.
[36] Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011;18(5):540–3.
[37] Resnik P, Niv M, Nossal M, Kapit A, Toern R. Communication of clinically relevant information in electronic health records: a comparison between structured data and unrestricted physician language. In: American health information management assoc. perspective in health information management, CAC proceedings; 2008.
[38] Kim JD, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. BMC Bioinform 2008;9:10.
[39] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Gene Ontol Consortium Nat Genet 2000;25(1):25–9.
[40] Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002;35(4):222–35.
[41] Tari L, Tu P, Hakenberg J, Chen Y, Son T, Gonzalez G, et al. Incremental information extraction using relational databases. IEEE Trans Known Eng 2012;24(1):86–99.
[42] Tari L, Anwar S, Liang S, Cai J, Baral C. Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics 2010;26(18):i547–53.
[43] Marcus M, Santorini B, Marcinkiewicz M. Building a large annotate corpus of English: the Penn Treebank. Comput Linguist 1993;19(2):313–30.
[44] Grishman R, Sundheim B. Design of the MUC-6 Evaluation. In: Sundheim B, editor. Proceedings of the fifth message understanding conference (MUC-5). San Mateo, CA., Morgan Kaufmann; 1996. p. 1–11.
[45] Humphreys K, Gaizauskas R, Azzam S, Huycs C, Mitchell B, Cunningham H, editors, et al. Proc of the seventh message understanding conference (MUC-7). Morgan; 1998.
[46] Yeniterzi R, Aberdeen J, Bayer S, Wellner B, Hirschman L, Malin B. Effects of personal identifier resynthesis on clinical text de-identification. J Am Med Inform Assoc 2010;17(2):159–68.
[47] Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc 2013;20(1):84–94.
[48] Fullerton SM, Wolf WA, Brothers KB, Clayton EW, Crawford DC, Denny JC, et al. Return of individual research results from genome-wide association studies: experience of the Electronic Medical Records and Genomics (eMERGE) Network. Genet Med 2012;14(4):424–31.
[49] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–13.

[50] Pustejovsky J. The syntax of event structure. Cognition 1991;41(1–3):47–81.
[51] Pustejovsky J, Castaño J, Ingria R, Sauri R, Gaizauskas R, Setzer A, et al. TimeML: robust specificatin of event and temporal expressions in text. In: Proc of 5th int workshop on computational semantic (IWCS-5); 2003.
[52] Jimino-Yepes A, Sticco J, Mork J, Aronson A. GeneRIF indexing: sentence selection based on machine learning. BMC Bioinform 2013 [in press].
[53] Cohen KB, Palmer M, Hunter L. Nominalization and alternations in biomedical language. PLoS One 2008;3(9):e3158.
[54] Sager N. Semantic formatting of scientific information. In: Walter de Gruyter, editor. AFIPS conf proc, vol. 41; 1972. p. 791–800.
[55] Hirschman L, Sager N. Automatic information formatting of a medical sublanguage. Sublanguage: studies of language in restricted domains. Walter de Gruyter; 1982. p. 27–80.
[56] Livingston K, Johnson H, Verspoor K, Hunter L. Leveraging gene ontology annotations to improve a memory-based language understanding system. In: Fourth IEEE int conf on semantic computing (IEEE ICSC2010); 2010.
[57] Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 2012;28(23):3158–60.
[58] Kilicoglu H, Fiszman M, Rodriquez A, Shin D, Ripple A, Rindflesch TC. Semantic MEDLINE: a web application to manage the results of PubMed searches. In: Proc third int symp for semantic mining in, biomedicine; 2008. p. 69–76.
[59] Rindflesch TC, Fiszman M, Rosemblat G, Shin D. Semantic MEDLINE: an advanced information management application for biomedicine. Inform Syst Use 2011;31(1 and 2):15–21.
[60] Wilkowski B, Fiszman M, Miller CM, Hristovski D, Arabandi S, Rosemblat G, et al. Graph-based methods for discovery browsing with semantic predications. In: AMIA annu symp proc; 2011. p. 1514–23.
[61] Miller CM, Rindflesch TC, Fiszman M, Hristovski D, Shin D, Rosemblat G, et al. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. Sleep 2012;35(2):279–85.
[62] Fiszman M, Bray BE, Shin D, Kilicoglu H, Bennett GC, Bodenreider O, et al. Combining relevance assignment with quality of the evidence to support guideline development. Stud Health Technol Inform 2010;160(Pt. 1):709–13.
[63] Miller RA, McNeil MA, Challinor SM, Masarie Jr FE, Myers JD. The INTERNIST-1/QUICK MEDICAL REFERENCE project–status report. West J Med 1986;145(6):816–22.
[64] Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. JAMA 1987;258(1):67–74.
[65] Warner Jr HR. Iliad: moving medical decision-making into new frontiers. Methods Inf Med 1989;28(4):370–2.
[66] Miller G. WordNet: a lexical database for English. Commun ACM 1995;38(11):39–41.